公募研究：2000～2001年度

タンパク質のアミノ酸間平均距離統計を用いたゲノム配列解析法の開発

●菊地武司

倉敷芸術科学大学産業科学技術学部（現在の所属：立命館大学情報理工学部）

〈研究の目的と進め方〉

ゲノム解析により様々な生物種のORF領域が明らかになっているが、個々のORFの立体構造を求め、その情報からタンパク機能などの情報を予測・抽出することが、その次のステップとして非常に重要である。しかしながら、複雑なORFの構造を一挙に解析することはなかなか困難である。したがってORF配列においてタンパクドメインをコードする領域を精度よく予測することは、立体構造や機能の予測を行う前のステップとして重要であるし、その後の過程にドメイン配列の情報を利用すれば効率のよい立体構造解析・立体構造予測が可能になると考えられる。そこで本研究では、ゲノム上のＯＲＦ配列における構造ドメインの位置の予測を目標とする。方法として、タンパク既知タンパクより計算したアミノ酸間平均距離のデータに基づき定義されたコンタクトマップ（Average Distance Map、ADM）をゲノムＯＲＦ配列に応用し、ドメインの位置の予測を試みる。その試みにおいて明らかになったADMの性質を分析し、その予測領域が何に対応しているのかを詳細に検討する。さらに、予測ドメイン領域の詳細な構造情報やフォールディング機構の予測についても検討を行う。最終的にドメイン予測ツールの開発を試みる。主な目標は、構造ドメインの位置の予測精度の向上と方法論の自動化である。

〈研究開始時の研究計画〉

当初の研究計画を以下に示す

① 国立遺伝研究所ゲノム構造予測データベースＧＴＯＰのデータを用いてE.coliと酵母のゲノム配列にADM法を適用し、構造ドメインの位置を予測する。その上で、ADMによる予測結果とGTOPによる予測結果（配列類似性に基づく）の比較を行い、ADMの予測能力を検証する。ドメインの位置の予測精度を向上させるため、アミノ酸組成や２次構造予測などの情報を組み合わせることも考慮する。

② ADMにより予測されている領域が実際どのような配列領域に対応しているかを詳細に検討する。そのため、フォールディング実験が成されているタンパクについて、ADMを応用し、フォールディングの性質と比較する。

③ 残基数の多いタンパクに対して、平均距離データの作成をやりなおすことと、タンパクを分割して、ドメインの位置を予測して最終的に双方の予測を組み合わせる。このようにして、サイズの大きいタンパクについてもＡＤＭを応用できるようにする。

④ 方法論の自動化を図る。

〈研究期間の成果〉

① 国立遺伝学研究所で開発されたゲノム構造予測データベースＧＴＯＰ中のＯＲＦに応用しＧＴＯＰ検索結果と比較した。その結果、ＡＤＭ法で予測したドメイン領域は、実際のコンタクトマップから解析した構造単位によく対応することがわかった。その予測精度は定性的な指標では70-80％になる。$\alpha\beta$型タンパクについて予測精度がよくないことが明確になったが、特定の$\alpha\beta$バレルタンパクについて予測を行うと、予測領域は機能単位に対応していることがわかった。

② 立体構造が類似しているが、フォールディング機構が異なると指摘されている3つのβサンドイッチ型タンパク質である脂肪酸結合タンパク、PDBコード１ＩＦＣ、１ＣＢＩ、１ＯＰＡに対し、ＡＤＭ法を適用し予測マップの差異を解析した。その結果、ＡＤＭで予測される構造のコア領域間の相互作用の違いがフォールディングの違いを反映することがわかった。同様の考察を植物ヘモグロビン、cタイプリゾチームに対して行い、ADMにおいて予測サブドメインの位置や性質の違いが、フォールディング機構の違いを反映することがわかった。このことはADMにより予測されるサブドメインの位置が、タンパクのフォールディング開始部位、構造形成単位に対応していることを示唆する。この知見は、精密はドメインの位置を予測する上でも重要な知見である。またコンタクトオーダー（ＣＯ）の計算に利用すると、立体構造に基づくＣＯとＡＤＭによる予測値の間に相関が見られることがわかった。このことはＡＤＭが構造形成速度に関する情報を含むことを示す。

③ 大きいサイズのORFに対し、あらかじめタンパクをいくつかのサイズに分割してADMによるドメインの位置の予測を行い、その予測を組み合わせることで大きいサイズのタンパクに対しても応用が可能であることが予備的に示された。より実用的な方法の基礎ができた。

④ 本方法の自動化を目標として、いくつかの自動化ルーチンの作成を行った。

〈国内外での成果の位置づけ〉

ドメインの位置の予測の試みは、アミノ酸配列相同性を用いる方法を除けば、さほど多くはない。その主な理由は、ドメインの定義がまだ明確ではないということによる。最近では、ドメインの位置のサイズ統計や、ドメイン境界のアミノ酸組成を用いた予測などが試みられているが、いずれにしても標準的方法には至っていない。本方法はこれまでのものと異なり、残基間平均距離統計から構造の核となる領域として予測を試みるので、ドメインの位置の予測が可能となる。本方法のオリジナルの論文（T. Kikuchi et al., J. Protein Chem., 7, 427, 1998 ）はドメインの位置の予測法の最初の試みの一つとして現在でもしばしば引用される。本研究期間において行われた研究にうち、特に海外からの反響があったものはβサンドイッチタンパクのフォールディング機構の差異に関する研究（上記「研究期間の成果」の②）であり、メキシコモレロス自治大学のDr. Arredondo-Peter (Laboratorio de Biofísicay Biología Molecular, Facultad de Ciencias, Universidad Autónoma del Estado de Morelos)から関心を

頂いた。Arredondo-Peter氏は植物ヘモグロビンの配列解析の専門家であり、新たな植物ヘモグロビン配列を見出している研究者である。氏が本研究の成果である論文(「研究期間の全成果公表リスト」4)-1)に関心を持ち、共同研究を我々に申し入れてきた。そして、ADM法を氏が研究している植物ヘモグロビン配列に応用し、植物ヘモグロビンの予測フォールディング単位と植物ヘモグロビンの進化の研究に発展し、一つの論文を纏めることができた。その成果が「研究期間の全成果公表リスト」4)-2のものである。

〈達成できなかったこと、予想外の困難、その理由〉

本研究期間において本研究課題に関するいくつかの知見が得られたことは上述の通りであるが、当初の目的に達成できなかったものや新たに明らかになった問題があることも事実である。

達成できなかった点とその理由を以下に示す。(番号は「研究開始時の研究計画」の番号に対応)

① ② α／βバレルタンパクのドメインの定義が予測の観点からは困難である。これは、このタイプのタンパクのフォールディング単位が一つではないことによる。従ってドメインの位置を精度よく予測するためには、まず構造型の判別を効果的に行う必要がある。これは、ゲノム配列のドメイン予測の問題において、パラドキシカルな関係にあり、やや困難な問題である。すなわち、ドメインの位置を予測するために、そこに含まれるそれぞれのORFの構造型を知る必要がある。そのためには、ORFの位置を予測しなければならないことになる。この点を明らかにするためには、 α／βバレル型タンパクのフォールディング機構とADMによる予測領域との関係を明確にする必要があり、予想以上に時間がかかることが判明した。

また、必ずしも機能ドメインと構造ドメインが一致しない。上述の問題とも関連しドメインの定義そのものにかかわる問題である。予測領域は、構造核あるいは構造ドメインに対応している。

③ ④ これらの問題の処理を先に行う必要があったため、大きいサイズのタンパクに対する方法論の改良や方法論の自動化を進めるのが遅れてしまった。

予想外の困難とその理由を以下に示す。(番号は「研究開始時の研究計画」の番号に対応)

① 上述の事項(① ②)と関連するが、フォールディング過程において、構造形成部位が1つでありその構造核が成長して構造をつくるような場合は、1ドメインタンパクとなる。この場合、ADMの予測はよく現実を反映する。構造形成部位が2つでありその構造核がそれぞれ独立に成長して構造をつくるような場合は、2ドメインタンパクとなりADMは明確にそのように予測する。しかしながら、複数の構造核(あるいは複数のフォールディングの遷移状態構造)が存在するが、それぞれが融合し一つのドメインを形成しているように見える場合が存在する。このときには、ADMは2つ以上のドメインを予測する。α／βバレルタンパクが典型的な例のように見える。これは、アミノ酸配列に対しいくつかの予測法を用意し、いくつかの予測結果を最終的に一つに絞り込むような作業を必要としていることを示唆している。この事態は実は当初は予想していなかったことであり、やや深刻な問題であると認識している。タンパクのフォールディング機構に関する統一的で深い理解を要求しているように思われる。

〈今後の課題〉

以上の問題点を克服するためには以下の課題に取り組まなければならないと考えている。

① ゲノムＯＲＦ配列への応用に適するように、方法論の自動化を進める。

② タンパクのフォールディング機構とADM予測領域との関係をさらに明確にし、ADMが何を予測しているか、すなわちADMにより何が予測可能かを明確化する。その上で、ドメインの位置の予測を再検討する。

③ ドメインの定義を再検討する。ドメインの今日的な定義を提案する。これまでは、ドメインの定義はその歴史的な経緯もあり、天然構造に立脚している。今後もこの定義は変わらないであろう。しかしながら、そのような定義においてもスーパーファミリー内のタンパクの構造を比較しても微妙な違いが存在しドメインの位置が(定義の仕方によって)変化する場合がある。様々なタンパクフォールディング機構が明らかになってくると、このような微妙な違いはフォールディング機構の差異を反映していることもありうる。従って、ドメインの定義もフォールディング機構(の進化的変化)を考慮したものである必要があるのかもしれない。その上でドメインの位置の予測法を再度検討する。

④ さらにそのような解析はタンパク構造形成とアミノ酸配列の関係の新たな知見をもたらすかもしれない。新たなアミノ酸配列比較法を提案する可能性があると思われる。この点においても、タンパク構造形成とアミノ酸配列に関する新しい定義の提出を試みる。基礎となる考え方はADM法、あるいは平均距離統計に基づくシミュレーションである。このような考え方は、新しいタンパク立体構造予測法に結びつくかもしれない。すなわち最終的には新しいタンパク立体構造予測法(精度の高い)の提案を視野に入れることができるのである。

〈研究期間の全成果公表リスト〉

1)論文／プロシーディング(査読付きのものに限る)

1. T. Kikuchi,: Contact Maps Derived from the Statistics of Average Distances between Residues in Proteins. Application to the Prediction of Structures and Active Sites of Proteins and Peptides. Recent Res. Devel. Protein Eng. 2, 1-48 (2002).)
「ゲノム特定オンラインニュース」登録受付番号:0111141122

2)データベース／ソフトウェア

得になし

3)特許など

得になし

4)その他

公募研究期間後の成果のため「ゲノム特定オンラインニュース」に登録していませんが、本研究課題と関連する成果として以下のものを挙げておきます。

1. T. Ichimaru and T. Kikuchi: Analysis of the Differences of the Folding Kinetics of Structurally Homologous Proteins based on the Predictions of the Gross feature of the Residue Contacts, PROTEINS, 51, 515-530 (2003).
2. Prediction of Folding Pathway and Kinetics Among Plant Hemoglobins by Using an Average Distance Map

Method.', Shunsuke Nakajima, Emma Álvarez-Salgado, Takeshi Kikuchi and Raúl Arredondo-Peter, PROTEINS, 61, 500-506 (2005).